Statistics and Its Interface Volume 12 (2019) 501–510

Statistical inference of the generalized Pareto distribution based on upper record values

Xu Zhao*, Xueyan Geng, Weihu Cheng, and Pengyue Zhang

Upper records are important statistics in environmental science and many other fields. Because upper records are crucial for policy making, precise modeling and inference techniques are in high demand. The generalized Pareto distribution (GPD) is commonly adopted by researchers for modeling heavy tail phenomena in many applications. The statistical inference of the GPD upper records is a critical issue in record analysis. Based on upper record data, the current parameter estimation methods of the GPD depend on preassumed shape parameter and only estimate the location and scale parameters. However, the shape parameter is typically unknown in real applications. In this manuscript, we propose a new approach that can estimate all three parameters of the GPD. The proposed estimator is used in conjunction with a moment method and nonlinear weighted least squares theory that minimizes the sum of squared deviations between the upper records and their expectations. In simulation studies, we compare alternative estimators and demonstrate that the new estimator is competitive in terms of the bias and means square error in estimating the shape and scale parameters. In addition, we investigate the performance of different threshold selection procedures by estimating the Value-at-Risk (VaR) of the GPD. Finally, we illustrate the utilization of the proposed methods by analyzing an air pollution data. In this analysis, we provide a detailed guide for selecting the threshold and upper records.

MSC 2010 subject classifications: 62F12.
Keywords and phrases: Generalized Pareto distribution, Extreme values, Upper record values, Parameter estimation, Threshold selection.

1. INTRODUCTION

Record values are defined as only observations greater (or less) than previous values. Such values are commonly seen in many fields, such as environmental science, sports and economics (Chandler (1952), Coles and Tawn (1996), Sultan, Al-Dayian and Mohammad (2008), Cramer and Naehrig (2012)). The importance of record values has been widely recognized by statisticians. Statistical research on record values started with Chandler (1952) and has now spread in many directions. To date, based on upper record values, studies have been conducted by Wang and Ye (2015) on the Weibull distribution, by Balakrishnan and Chan (1998) and Shahab, Al-Dayian and El-Beltagi (2001) on the normal distribution, by Raqab (2002) on the generalized exponential distribution and by Soliman et al. (2010) on the inverse Rayleigh distribution. Although these studies provide guidance for modeling record values, the data often deviate from the distribution assumptions. The GPD is known as a powerful tool for addressing extreme values and has great potential for modeling record values (Ahsanullah (2004)). Pickands (1975) and Balkema and De Haan (1974) noted that exceedances over a high threshold can be asymptotically fitted by the GPD if the distribution of the complete sample is in the maximum domain of attraction, known as the peaks-over-threshold (POT) method in extreme value theory (EVT). In fact, most of the common continuous distributions meet this condition (Embrechts, Klüppelberg and Mikosch (1997)). Under the POT framework, we can apply the GPD to fit the tail region of a data set, even if we do not know the underlying distribution. Therefore, the statistical inference of the GPD upper records is a critical issue in record analysis. For various quantitative risk analyses and the prediction of future upper records, two fundamental steps in the application of the GPD are parameter estimation and threshold selection.

Research is ongoing for the two above mentioned steps. First, the parameter estimation of the GPD must be based on a prespecified shape parameter. For instance, using upper record values, Ahsanullah (2004) derived the best linear unbiased estimator (BLUE) and best linear invariant estimator (BLIE) for the location and scale parameters. Notably, Sultan and Moshref (2000) provided approximate confidence intervals based on the BLUE for the location and scale parameters. Second, if the GPD record value series is unknown and selected in advance, the key step is finding an optimal threshold as the first record value, such that exceedances over the threshold follow the GPD separately. On one hand, if the threshold is below the optimal value, the GPD assumption may be violated. On the other hand, the variation may be inflated if the threshold is above the optimal value. Graphical diagnosis methods have been widely applied for threshold estimation (Davison and Smith (1990), Dress, De Haan and Resnick (2000), Coles (2001), Scarrott and MacDonald (2012)). Alternatively, threshold can be selected using goodness-of-fit test or automatic sequential goodness-of-fit tests. These practical approaches yield

*Corresponding author.

satisfactory performance. Recently, Bader, Yan and Zhang (2018) proposed an automated threshold selection procedure based on a sequence of goodness-of-fit tests and the stopping rule in G'Sell et al. (2016). This approach is elective and automated and can be used to obtain the GPD record values for statistical inference.

In real applications, however, the GPD shape parameter is generally unknown, and in many cases the GPD record value series is unknown and must be selected in advance. Because of these difficulties, we propose a new GPD estimator and guide for selecting upper records. Our method is based on a moment method and nonlinear weighted least squares theory. The proposed estimator addresses a caveat of shape parameter estimation for the GPD based on upper records. The contributions of the current article are as follows. First, we use the nonlinear weighted least squares moment estimated equations to estimate the shape and scale parameters simultaneously and present an optimization procedure. Thus, the GPD can be applied without special assumptions regarding the shape parameter. Second, we illustrate how to select the optimal threshold above which the GPD fits the exceedances. The chosen threshold can be regarded as the first record for location parameter estimation, and the corresponding upper records can be determined. Using various simulations, we compare the performance of four estimators and show that the proposed weighted nonlinear least squares moment (WNLSM) estimator is competitive compared to other estimators. Using a realistic Beijing PM2.5 dataset, we apply three threshold selection methods to determine the optimal threshold and choose upper records. Following the proposed estimation approach, we estimate the GPD parameters and predict future upper records. These results serve as references for environmental agencies and Beijing residents.

The remainder of this paper is organized as follows. In section 2, exact expressions for the single moments of the upper record values are proposed. Based on upper records, in section 3, we describe a new GPD parameter estimator and introduce a recently developed threshold selection technique to find the first record value. In section 4, we compare the performance of different parameter estimation approaches based on the GPD with simulated and real data sets, as well as the performance of different optimal threshold methods for mixture distributions with various simulation studies. The conclusions are presented in section 5.

2. MOMENTS OF UPPER RECORD DATA

In this section, notations and definitions are given in subsection 2.1. In subsection 2.2, we derive the exact explicit expressions for the single moments of the upper record values from the GPD.

2.1 Notations and definitions

Let $\{X_n, n = 1, 2, \cdots\}$ be an infinite sequence of independent and identical random variables with the same distribution as the random variable X. Denote the cumulative distribution function (cdf) of X as F and the probability distribution function (pdf) of X as f. $X_{L(1)}, X_{L(2)}, \cdots$ are called lower record value statistics if $L(i) = \min\{j : j > L(i-1), X_j < X_{L(i-1)}\}$, where $i \geq 2$ with $L(1) = 1$. An analogous definition can be given for upper record values. Let $X_{U(1)}, X_{U(2)}, \cdots$ denote upper record values, where $U(i) = \min\{j : j > U(i-1), X_j > X_{U(i-1)}\}$, and $i \geq 2$ with $U(1) = 1$. The pdf of the lower record $X_{L(i)} (i = 1, 2, \cdots)$ is given by (see Arnold, Balakrishnan and Nagaraja (1998))

$$f_{L(i)}(x) = \frac{1}{\Gamma(i)}[-\ln F(x)]^{i-1} f(x)$$

and the pdf of upper record $X_{U(i)} (i = 1, 2, \cdots)$ is given by (see Arnold, Balakrishnan and Nagaraja (1998))

$$f_{U(i)}(x) = \frac{1}{\Gamma(i)}[-\ln(1 - F(x))]^{i-1} f(x)$$

where $\Gamma(x)$ is a gamma function.

2.2 Moments of the GPD upper record data

The cdf of the GPD(μ, σ, ξ) with the location parameter μ, the scale parameter σ and the shape parameter ξ, respectively, is defined as follows.

$$(1) \qquad G_{\mu,\sigma,\xi}(x) = \begin{cases} 1 - \left(1 + \xi \dfrac{x-\mu}{\sigma}\right)^{-1/\xi}, & \xi \neq 0 \\ 1 - \exp\left(-\dfrac{x-\mu}{\sigma}\right), & \xi = 0 \end{cases}$$

If $\xi \geq 0$,$x \geq \mu$, else $\mu < x \leq \mu - \sigma/\xi$. The exponential distribution is obtained taking $\xi = 0$ in (1).

The pdf of the GPD is given by

$$g_{\mu,\sigma,\xi}(x) = \begin{cases} \dfrac{1}{\sigma}\left(1 + \xi \dfrac{x-\mu}{\sigma}\right)^{-1/\xi-1}, & \xi \neq 0 \\ \dfrac{1}{\sigma}\exp\left(-\dfrac{x-\mu}{\sigma}\right), & \xi = 0 \end{cases}$$

Let $X_{U(1)}, X_{U(2)}, \cdots, X_{U(n)}$ be a sequence of n upper record values from the three-parameter GPD. It is easy to prove that,

$$(2) \qquad \frac{X_{U(i)} - \mu}{\sigma} = \frac{1}{\xi}[(Y_{L(i)})^{-\xi} - 1]$$

where $Y_{L(1)}, Y_{L(2)}, \cdots, Y_{L(n)}$ is a sequence of n lower record values from the uniform distribution U(0,1). Next, we obtain the single moments of the GPD upper record data using the above relation in equation (2).

Theorem 2.1. *For a fixed positive integer $r \geq 1$, the single moments $\mathrm{E}(Y_{L(i)}^{-\xi})^r$ is given as*

$$(3) \qquad \mathrm{E}(Y_{L(i)}^{-\xi})^r = (1 - \xi r)^{-i}, \qquad \xi < 1/r.$$

Especially, as a check put $r = 1$ in (3) we get

$$\mathrm{E}(X_{U(i)}) = \mu - b + b(1-\xi)^{-i}, \qquad \xi < 1 \tag{4}$$

where $b = \sigma/\xi$. A simplified method of computing the GPD single moment is given in equation (4) using the U(0,1). The result is equivalent to the findings of Balakrishnan and Ahsanullah (1994), Sultan and Moshref (2000) and Ahsanullah (2004).

3. GPD PARAMETER ESTIMATION

The survival function of a continuous random variable X is denoted as $\bar{F}(x) = 1 - F(x), 0 < x < \infty$. Then $F(x)$ is regularly varying with index$-\xi < 0$, or simply $\bar{F} \in \mathfrak{R}_{-\xi}$, if

$$\lim_{x\to\infty} \frac{\bar{F}(x\lambda)}{\bar{F}(x)} = \lambda^{-\xi}, \lambda > 0.$$

Heavy-tailed distributions such as the Pareto, generalized Pareto, Log-gamma, Cauchy and Stable distributions are examples of such functions (Park and Kim (2016)).

A principle result of EVT is the famous Pickands-Balkema-de Haan theorem (Balkema and De Haan (1974), Pickands (1975)) which states that, for $\bar{F} \in \mathfrak{R}_{-\xi}$, the excess loss $(X - u|X > u)$ from such a distribution F with a large threshold $u > 0$ converges to the two-parameter GPD with $\xi > 0$. That is

$$\lim_{u\uparrow x_F} \sup_{0\le y<x_F-u} |F_u(y) - G_{\sigma,\xi}(y)| = 0$$

where $F_u(y) = P(X - u \le y|X > u)$ with the support $0 \le y < x_F - u$. $x_F \le \infty$ is the right endpoint of F and $G_{\sigma,\xi}(\cdot)$ is the GPD distribution function.

This relation implies that the excess distribution $F_u(y)$ converges to the two-parameter GPD when the threshold u is appropriately selected whenever X is a heavy-tailed distribution. Therefore, the location parameter μ can be estimated by u, because $G_{u,\sigma,\xi}(x) = G_{\sigma,\xi}(x - u)$ with the support set $\{X|X > u\}$.

3.1 Threshold selection

Before estimating parameters, one important step is determining the optimal threshold u. If the selected threshold is too low, the GPD approximation may be not satisfied. In contrast, if the threshold is too high, the small sample size may increase the variance of parameter estimation. Our solution is to follow a recently developed stopping rule (Bader, Yan and Zhang (2018)). This approach combines the sequential goodness-of-fit testing proposed by Choulakian and Stephens (2001) and the stopping rule of G'Sell et al. (2016).

Consider a given sequence of candidate thresholds $u_1 < \cdots < u_m$, where m can be fixed at $m = n - 10$ or $n - 20$ (Langousis et al. (2016)) and n is the sample size. For each u_i $(i = 1, \cdots, m)$, there is a corresponding null hypothesis given by

$$H_0^{(i)} : \text{The excess loss } (X - u_i|X > u_i) \text{ follows the GPD.} \tag{5}$$

The ForwardStop rule of G'Sell et al. (2016) is given by

$$\hat{k} = \max\left\{k \in \{1, \cdots, m\} : -\frac{1}{k}\sum_{i=1}^{k} \log(1 - p_i) \le \alpha\right\} \tag{6}$$

where α is a prespecified level and $p_1, \cdots, p_m$ are the corresponding p-values of the m hypotheses. The rejection rule is constructed by returning a cutoff $\hat{k}$ such that $H_0^{(1)}, \cdots, H_0^{(\hat{k})}$ are rejected (Bader, Yan and Zhang (2018)). This testing method involves m times goodness-of-fit tests, as noted by Choulakian and Stephens (2001), even if $H_0^{(i)}$ is accepted. The reason for conducting multiple tests is that unless the test has high power, $H_0^{(i)}$ may be accepted at a low threshold by chance. In such a case, the GPD may not fit all the exceedances above the chosen threshold.

3.2 Weighted nonlinear least squares moment estimation method

The selected u can be regarded as the first record $X_{U(1)}$ and the upper records $X_{U(1)}, \cdots, X_{U(n)}$ are recorded from the original sequence $\{X_1, X_2, \cdots\}$. Under the POT framework, the excess loss $(X - u|X > u)$ converges to the two-parameter GPD which has only shape and scale parameters. Therefore, we are only interested in estimating the shape and scale parameters.

Now, we introduce a new estimator for the GPD by minimizing the sum of squared deviations between the upper records and their corresponding expectations. That is

$$(\hat{\xi}, \hat{b}) = \arg\min_{(\xi,b)} \sum_{i=1}^{n} \left[X_{U(i)} - \mathrm{E}(X_{U(i)})\right]^2. \tag{7}$$

Then $\hat{\sigma} = \hat{\xi}\hat{b}$ and $(\hat{\xi}, \hat{\sigma})$ is the least square estimator (LSE). However, a direct fitting approach does not work well, because the expectation of the upper record values is sensitive to the shape parameter ξ, as shown in equation (4). In view of this problem, we use a nonlinear least squares (NLS) method with a three-step fitting procedure to find stable estimators.

The first step is to find the interim estimate using a nonlinear minimization of upper records:

$$(\hat{\xi}, \hat{b}) = \arg\min_{(\xi,b)} \sum_{i=1}^{n} \left\{\exp(X_{U(i)}) - \exp[\mathrm{E}(X_{U(i)})]\right\}^2. \tag{8}$$

Before solving optimization equation (8), we can use the concept of the moment estimation method and construct a moment estimation equation based on the expectation of

the upper records in equation (4). The moment estimation equation is given as follows.

$$\bar{X}_U = \frac{1}{n}\sum_{i=1}^{n} \mathrm{E}(X_{U(i)}) = b[A(\xi) - 1], \quad \xi < 1 \tag{9}$$

where $\bar{X}_U = \frac{1}{n}\sum_{i=1}^{n} X_{U(i)}, A(\xi) = \frac{1}{n}\sum_{i=1}^{n}(1-\xi)^{-i}$. When rearranged, equation (9) can be written as

$$b = \frac{\bar{X}_U}{A(\xi) - 1}. \tag{10}$$

Replacing the b in equation (8), the above optimization equation (8) is written as:

$$\hat{\xi_1} = \arg\min_{\xi} \sum_{i=1}^{n} \left\{ \exp(X_{U(i)}) - \exp\left[\frac{\bar{X}_U}{A(\xi) - 1}\left((1-\xi)^{-i} - 1\right)\right]\right\}^2. \tag{11}$$

From equation (10), $\hat{\sigma_1} = \dfrac{\hat{\xi_1}\bar{X}_U}{A(\hat{\xi_1}) - 1}$.

Compared the equation (11) with the equation (8), this setup equation (11) is advantageous in that the b has been eliminated. Therefore, the shape parameter ξ can be separately estimated and is independent of b, making this method more efficient. The result of the first step $(\hat{\xi_1}, \hat{\sigma_1})$ is called the nonlinear least squares moment 1 (NLSM1) estimator.

Second, with $\hat{\xi_1}$ as the initial value, the following step includes the optimization according to the upper records and their expectations.

$$\begin{aligned}\hat{\xi_2} &= \arg\min_{\xi} \sum_{i=1}^{n} [X_{U(i)} - \mathrm{E}(X_{U(i)})]^2 \\ &= \arg\min_{\xi} \sum_{i=1}^{n} \left[X_{U(i)} - \frac{\bar{X}_U}{A(\xi) - 1}\left((1-\xi)^{-i} - 1\right)\right]^2.\end{aligned} \tag{12}$$

Then $\hat{\sigma_2} = \dfrac{\hat{\xi_2}\bar{X}_U}{A(\hat{\xi_2}) - 1}$. $(\hat{\xi_2}, \hat{\sigma_2})$ is called the nonlinear least squares moment 2 (NLSM2) estimator.

Third, with $\hat{\xi_2}$ as the initial value and recognizing that $X_{U(i)}$ has different standard variances for various i values, we extend the NLSM2 estimator to the weighted NLSM (WNLSM) version by minimizing the following equation.

$$\begin{aligned}\hat{\xi_3} &= \arg\min_{\xi} \sum_{i=1}^{n} \omega_i [X_{U(i)} - \mathrm{E}(X_{U(i)})]^2 \\ &= \arg\min_{\xi} \sum_{i=1}^{n} \omega_i \left[X_{U(i)} - \frac{\bar{X}_U}{A(\xi) - 1}\left((1-\xi)^{-i} - 1\right)\right]^2\end{aligned}$$

where

$$\begin{aligned}\omega_i &= [\mathrm{Var}(X_{U(i)})]^{-1/2} \\ &= \frac{A(\xi) - 1}{\bar{X}_U}\left[\frac{1}{(1-2\xi)^i} - \frac{1}{(1-\xi)^{2i}}\right]^{-1/2}, \quad \xi < 1/2.\end{aligned}$$

And $\hat{\sigma_3} = \dfrac{\hat{\xi_3}\bar{X}_U}{A(\hat{\xi_3}) - 1}$. We will call $(\hat{\xi_3}, \hat{\sigma_3})$ the WNLSM estimator. One advantage of the WNLSM estimator over the NLSM1 and 2, as will be seen in later simulation, is that it estimates parameters in a more stable manner because, as $X_{U(i)}$ moves towards the tail side, the weight ω_i becomes smaller. We apply the "optim" function in R to solve optimizations, see Nelder and Mead (1965).

4. NUMERICAL ILLUSTRATION

In this section, we first conduct extensive simulation studies to compare alternative estimators and show that the proposed WNLSM estimator is highly competitive. Second, we investigate the performance of three threshold selection methods under various parametric mixture distributions. Third, we use daily PM2.5 record data from Beijing to illustrate the utility of the proposed methods in detail and identify important air pollution trends.

4.1 Simulation study of the parameter estimation

In this simulation study, we estimate parameters only using the proposed LS, NLSM1, NLSM2 and WNLSM estimators from various parametric GPD distributions. We do not compare the existing estimation approaches cited in section 1, because these methods provide the estimators of the GPD location and scale parameters and assume that the shape parameter is known. We describe the simulation procedure as follows.

1. Generate *iid* observations following the GPD.
2. Select the first observation as the first upper record $X_{U(1)}$.
3. Choose the upper record sequence $X_{U(1)}, \cdots, X_{U(n)}$ from the observations.
4. Use the proposed LS, NLSM1, NLSM2 and WNLSM estimators to estimate (σ, ξ).

Conduct 1000 simulations under different conditions to evaluate the mean square error (MSE) and Bias. We consider four different parameter pairs for parameter estimation based on the GPD: $(\mu, \sigma, \xi) = (0, 1, \pm 0.1)$ and $(0, 1, \pm 0.4)$, and records sample sizes $n = 3$ to 7. It should be noted that to obtain a fixed number of upper records, the size of the complete sample differs in each iteration. As shown in Table 1, the WNLSM estimator performs best for most parameter choices in terms of both the MSE and Bias.

Table 1. Parameter estimation under the GPD based on the upper records

n	(σ,ξ)	Method	MSE		Bias	
			σ	ξ	σ	ξ
3						
	(1,-0.4)	LS	6.246	1.761	2.499	-1.327
		NLSM1	5.833	1.649	2.415	-1.284
		NLSM2	6.231	1.760	2.496	-1.327
		WNLSM	2.110	0.462	1.453	-0.680
	(1,-0.1)	LS	2.971	0.366	1.724	-0.605
		NLSM1	2.396	0.301	1.548	-0.549
		NLSM2	2.945	0.364	1.716	-0.604
		WNLSM	1.001	0.075	1.000	-0.273
	(1,0.1)	LS	9.260	0.514	3.043	-0.717
		NLSM1	5.571	0.357	2.360	-0.597
		NLSM2	6.377	0.401	2.525	-0.633
		WNLSM	2.574	0.127	1.604	-0.357
	(1,0.4)	LS	4.297	0.168	2.072	-0.410
		NLSM1	3.373	0.151	1.837	-0.389
		NLSM2	4.205	0.167	2.051	-0.409
		WNLSM	2.333	0.096	1.528	-0.311
4						
	(1,-0.4)	LS	2.055	0.468	1.434	-0.684
		NLSM1	1.779	0.396	1.334	-0.630
		NLSM2	2.042	0.465	1.429	-0.682
		WNLSM	0.446	0.044	0.668	-0.209
	(1,-0.1)	LS	0.818	0.054	0.904	-0.232
		NLSM1	0.493	0.029	0.702	-0.171
		NLSM2	0.810	0.053	0.900	-0.231
		WNLSM	0.162	0.000	0.403	-0.010
	(1,0.1)	LS	0.414	0.019	0.643	-0.139
		NLSM1	0.219	0.011	0.468	-0.107
		NLSM2	0.401	0.019	0.633	-0.138
		WNLSM	0.032	0.000	0.180	0.018
	(1,0.4)	LS	2.055	0.076	1.434	-0.275
		NLSM1	4.126	0.084	2.031	-0.290
		NLSM2	1.934	0.076	1.391	-0.275
		WNLSM	0.636	0.021	0.797	-0.145
5						
	(1,-0.4)	LS	1.381	0.249	1.175	-0.499
		NLSM1	1.140	0.200	1.067	-0.448
		NLSM2	1.367	0.247	1.169	-0.497
		WNLSM	0.272	0.023	0.522	-0.152
	(1,-0.1)	LS	0.403	0.014	0.634	-0.119
		NLSM1	0.228	0.006	0.477	-0.079
		NLSM2	0.393	0.014	0.627	-0.117
		WNLSM	0.072	0.001	0.268	0.026
	(1,0.1)	LS	0.251	0.004	0.501	-0.062
		NLSM1	0.166	0.003	0.407	-0.051
		NLSM2	0.244	0.004	0.494	-0.062
		WNLSM	0.043	0.001	0.208	0.032
	(1,0.4)	LS	0.964	0.024	0.982	-0.153
		NLSM1	4.326	0.048	2.080	-0.218
		NLSM2	0.832	0.023	0.912	-0.153
		WNLSM	0.226	0.008	0.475	-0.088

Table 1. (Continued)

n	(σ,ξ)	Method	MSE		Bias	
			σ	ξ	σ	ξ
6						
	(1,-0.4)	LS	1.862	0.331	1.365	-0.576
		NLSM1	1.539	0.270	1.241	-0.519
		NLSM2	1.849	0.329	1.360	-0.574
		WNLSM	0.291	0.031	0.539	-0.176
	(1,-0.1)	LS	0.150	0.004	0.387	-0.066
		NLSM1	0.088	1.83e-03	0.296	-0.043
		NLSM2	0.143	0.004	0.379	-0.064
		WNLSM	0.025	1.66e-03	0.158	0.041
	(1,0.1)	LS	0.126	2.58e-03	0.355	-0.051
		NLSM1	0.104	2.23e-03	0.323	-0.047
		NLSM2	0.118	2.55e-03	0.344	-0.051
		WNLSM	0.013	1.51e-03	0.113	0.039
	(1,0.4)	LS	0.973	0.013	0.986	-0.112
		NLSM1	100.5	0.259	10.03	-0.509
		NLSM2	0.835	0.013	0.914	-0.114
		WNLSM	0.164	0.004	0.405	-0.059
7						
	(1,-0.4)	LS	1.129	0.192	1.062	-0.439
		NLSM1	0.882	0.147	0.939	-0.383
		NLSM2	1.122	0.191	1.059	-0.437
		WNLSM	0.174	0.017	0.417	-0.129
	(1,-0.1)	LS	0.089	0.001	0.299	-0.037
		NLSM1	0.052	0.000	0.228	-0.022
		NLSM2	0.086	0.001	0.293	-0.036
		WNLSM	0.020	0.001	0.140	0.032
	(1,0.1)	LS	0.072	9.4e-04	0.269	-0.031
		NLSM1	0.053	7.2e-04	0.231	-0.027
		NLSM2	0.068	9.3e-04	0.261	-0.031
		WNLSM	0.004	1.9e-03	0.066	0.043
	(1,0.4)	LS	0.657	0.013	0.811	0.112
		NLSM1	103.5	0.209	10.17	-0.457
		NLSM2	0.544	0.013	0.738	-0.115
		WNLSM	0.012	0.002	0.111	-0.041

To comprehensively compare the performance of different estimation methods, we consider ξ in a range, and (μ,σ)=(0,1) for record sample size $n = 3$. Figures 1 and 2 display the Bias and MSE for σ and ξ respectively. The Bias and MSE were estimated from 1,000 simulations. Overall we see that the proposed WNLSM is always superior to other methods for both scale and shape parameters. Thus the WNLSM estimator generally improves the estimation quality.

4.2 Simulation study of the threshold selection

We compare the performance of different threshold selection procedures by estimating the VaR which is the $100p$ quantile of the GPD, denoted by

$$\text{VaR}_p = u + \frac{\sigma}{\xi}\left[\left(\frac{F_0}{1-p}\right)^{\xi} - 1\right] \tag{13}$$

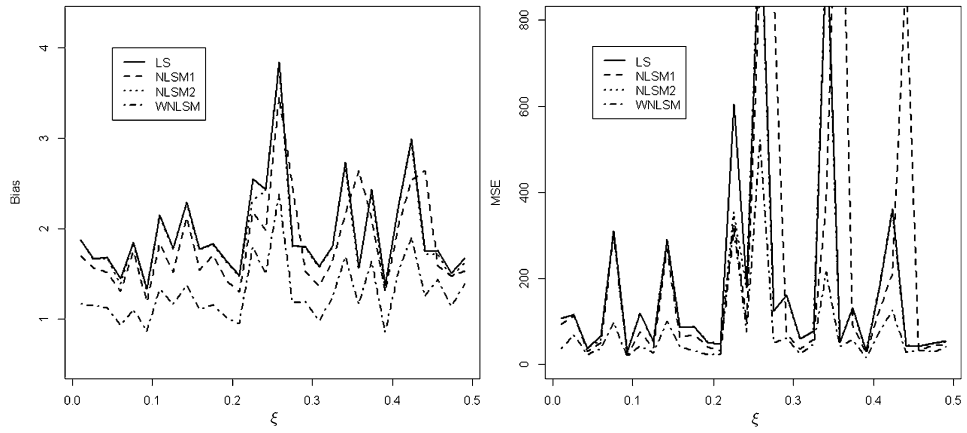

Figure 1. σ *estimation under the GPD when* $(\mu, \sigma) = (0, 1), \xi \in (0, 0.5)$ *and* $n = 3$.

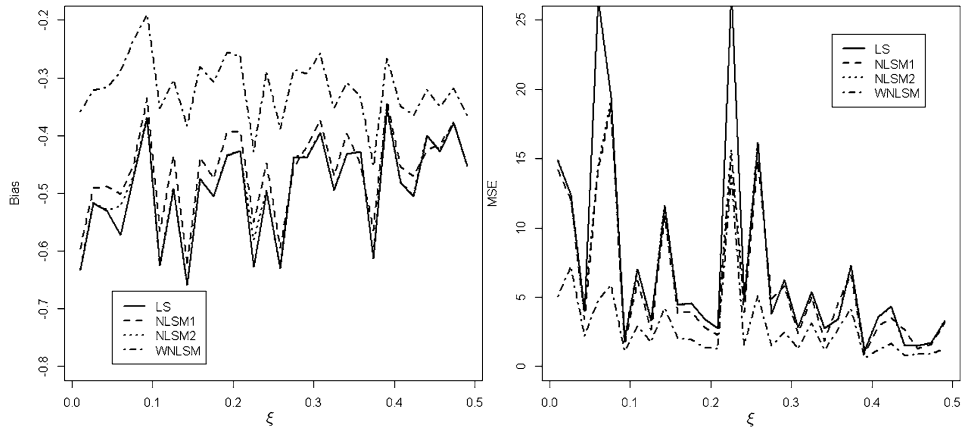

Figure 2. ξ *estimation under the GPD when* $(\mu, \sigma) = (0, 1), \xi \in (0, 0.5)$ *and* $n = 3$.

for a given threshold u, where F_0 is the proportion of the data exceeding u. The VaR is widely used to qualify tail risk in several fields, with p close to 1.

In addition to ForwardStop, two competing ascending and descending goodness-of-fit tests in tail risk measures VaR were used for comparison. These two tests can be implemented two ways to ForwardStop procedure (Bader, Yan and Zhang (2018)). Consider a fixed sequence of sorted order candidate thresholds. Ascending goodness-of-fit test starts from the first threshold and chooses the lowest threshold until an acceptance of $H_0^{(i)}$ as given in equation (5) happens. Descending goodness-of-fit test begins at the last threshold and descends until a rejection of the test occurs.

The simulation samples are generated from a mixture distribution. Mixture distribution consists of a 50/50 split from a distribution and GPD. It implies that the drawn probabilities of a single observation from the distribution and GPD are all 0.5. We consider different parameter pairs for the VaR estimation under the mixture distributions: 50/50 mixture of Beta: $(a, b) = (2, 1)$ and GPD: $(\sigma, \xi) = (0.5, 0.25)$ with $u = 1$, $(\sigma, \xi) = (1, 0.5)$ with $u = 2$, $(\sigma, \xi) = (2.5, 0.25)$ with $u = 5$. 50/50 mixture of Weibull: $(\kappa, \beta) = (0.45, 1)$ and GPD: $(\sigma, \xi) = (1, 0.4)$ with $u = 0.4428726$, $(\sigma, \xi) = (1.419845, 0.2)$ with $u = 0.4428726$. Our selections imply that $u = 0.4428726$ corresponds to the 50.0th percentile of the Weibull distribution. The cdf and pdf for these five parameter pairs mixture distributions can be seen in Figures 3 and 4.

In order to compare the performance of various threshold selection procedures, we set up the Monte Carlo simulation as follows.

1. Generate a random sample $\{X_1, \cdots, X_n\}$ of size $n = 200$ from the given mixture distribution, and transform to the order statistics $\{X_{1:n}, \cdots, X_{n:n}\}$.

2. Select the optimal threshold by below three methods applying R package "eva" (Bader and Yan (2015)).

a). ForwardStop. In this step, $u_{\hat{k}} = X_{\hat{k}:n}$, where $\hat{k}$ is given in equation (6).

b). Ascending goodness-of-fit test. Begin with $X_{i:n}$ where $i = 1$ and continue for $i = i+1$ until the GPD null hypothesis $H_0^{(i)}$ as given in (5) is accepted. If all are rejected, choose threshold $u = X_{m:n}$ where m as given in subsection 3.1.

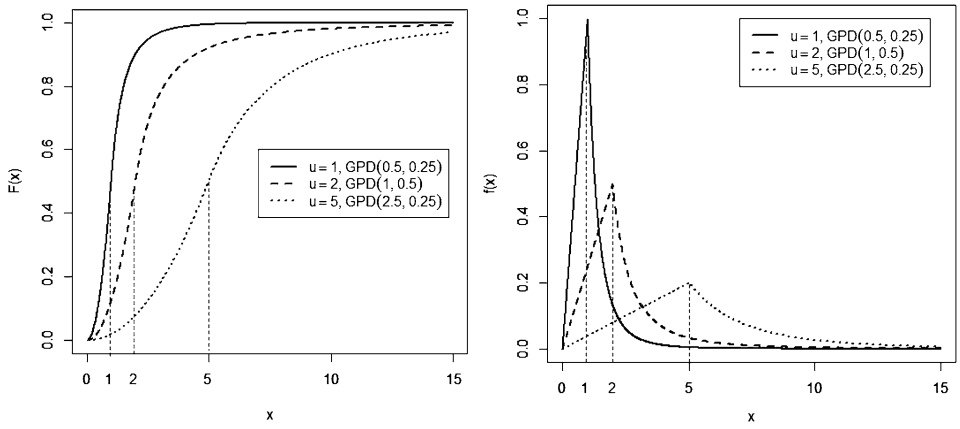

Figure 3. The distribution function $F(x)$ and probability density function $f(x)$ for the 50/50 mixture of the Beta: $(a, b) = (2, 1)$ and the GPD with three parameter pairs.

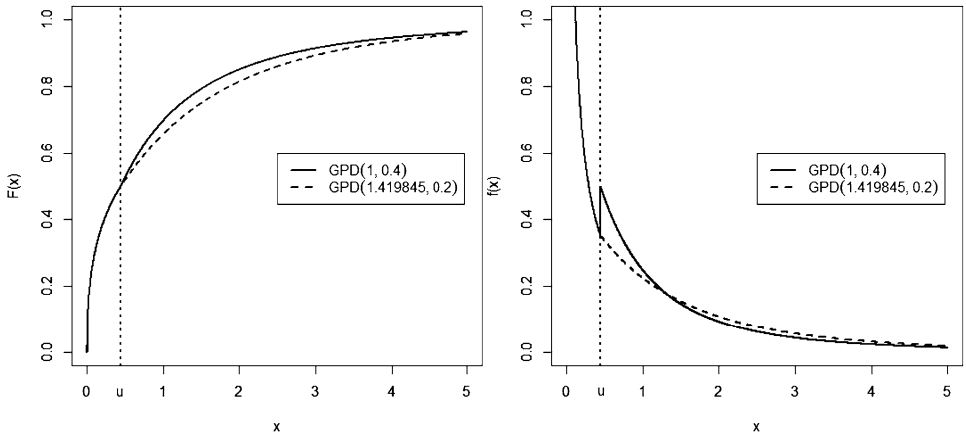

Figure 4. The distribution function $F(x)$ and probability density function $f(x)$ for the 50/50 mixture of the Weibull: $(\kappa, \beta) = (0.45, 1)$ and the GPD with two parameter pairs. The change-point at $u = 0.4428726$.

c). Descending goodness-of-fit test. Begin with $X_{i:n}$ where $i = m$ and continue for $i = i - 1$ until $H_0^{(i)}$ is rejected.

3. Fit the GPD with the excess $X - u|X > u$ to obtain the GPD parameter estimators $\hat{\xi}$ and $\hat{\sigma}$, which are used to estimate VaR given in equation (13).

4. Repeat above steps 1,000 times to compute the Bias and MSE of each VaR estimator.

Table 2 presents the MSE and Bias of the VaR 95%, 98% and 99%. Overall we see that there is no clear-cut winner in estimating the VaR. ForwardStop is, however, more stable and less sensitive to ascending and descending tests. First, in terms of the MSE, on one hand, based on the Beta and GPD mixture distributions, ForwardStop and ascending procedures perform best almost across all VaR extreme levels considered and all tested significance levels. On the other hand, based on the Weibull and GPD mixture distributions, ForwardStop and descending tests form a better result for VaR 98% and 99%, but ForwardStop and ascending methods perform better for VaR 95% under all conditions. Second, in term of the Bias, based on all mixture distributions, no approach outperforms in all setting. For example, when significance level $\alpha = 0.05$, ForwardStop gives the smallest Bias for VaR 95% and 98% based on the Beta and GPD mixture distributions with one exception, and for all VaR extreme levels considered under the Weibull and GPD mixture distributions.

4.3 Beijing PM 2.5 data analysis

We analyze recent daily PM2.5 data from Beijing collected in autumn and winter, from September 1, 2016 to February 28, 2017 and September 1, 2017 to February 28, 2018, because air pollution is most serious in these two seasons in Beijing. The sample size is 361. We focus on the regularity and trend of extreme pollution conditions to establish an appropriate model for forecasting the record-breaking degree of contamination. Then, an effective risk protection program can be established so that people can quickly respond to extreme pollution events.

Table 2. VaR estimation under the mixture distribution

α	Method	MSE			Bias		
		VaR 95%	VaR 98%	VaR 99%	VaR 95%	VaR 98%	VaR 99%
50/50 mixture of Beta(2,1) and GPD(0.5,0.25) with $u = 1$							
	ForwardStop	0.046	0.180	0.504	0.023	-0.100	-0.266
0.01	Ascending	0.047	0.176	0.499	0.036	-0.119	-0.336
	Descending	0.057	0.195	0.584	0.001	-0.083	-0.166
	ForwardStop	0.046	0.193	0.550	-0.004	-0.076	-0.161
0.05	Ascending	0.045	0.182	0.510	0.012	-0.112	-0.283
	Descending	0.059	0.210	0.626	-0.027	-0.093	-0.135
	ForwardStop	0.048	0.198	0.590	-0.011	-0.066	-0.121
0.1	Ascending	0.045	0.186	0.516	0.001	-0.106	-0.249
	Descending	0.061	0.213	0.618	-0.028	-0.097	-0.143
50/50 mixture of Beta(2,1) and GPD(1,0.5) with $u = 2$							
	ForwardStop	0.551	3.507	13.2	0.018	-0.464	-1.228
0.01	Ascending	0.542	3.328	12.5	0.044	-0.635	-1.752
	Descending	0.675	3.869	16.7	0.014	-0.281	-0.636
	ForwardStop	0.580	3.922	15.9	-0.016	-0.238	-0.504
0.05	Ascending	0.548	3.464	12.8	0.001	-0.511	-1.332
	Descending	0.752	4.180	18.8	-0.057	-0.220	-0.212
	ForwardStop	0.632	4.167	18.1	-0.026	-0.144	-0.168
0.1	Ascending	0.551	3.618	13.7	-0.016	-0.427	-1.064
	Descending	0.775	4.313	20.6	-0.063	-0.231	-0.183
50/50 mixture of Beta(2,1) and GPD(2.5,0.25) with $u = 5$							
	ForwardStop	1.166	4.385	12.2	0.162	-0.440	-1.272
0.01	Ascending	1.193	4.280	12.1	0.224	-0.543	-1.629
	Descending	1.482	4.948	14.9	0.072	-0.321	-0.718
	ForwardStop	1.204	4.743	13.2	0.038	-0.318	-0.765
0.05	Ascending	1.141	4.443	12.3	0.101	-0.509	-1.357
	Descending	1.690	5.274	14.7	-0.045	-0.388	-0.667
	ForwardStop	1.282	4.895	13.8	0.004	-0.276	-0.575
0.1	Ascending	1.146	4.515	12.4	0.051	-0.476	-1.196
	Descending	1.707	5.267	14.6	-0.042	-0.409	-0.722
50/50 mixture of Weibull(0.45, 1) and GPD(1, 0.4) with $u = 0.4428726$							
	ForwardStop	0.396	2.631	9.813	0.034	0.135	0.334
0.01	Ascending	0.433	3.439	14.1	0.067	0.438	1.076
	Descending	0.469	2.344	8.525	0.004	-0.003	0.092
	ForwardStop	0.398	2.370	8.903	-0.003	-0.016	0.043
0.05	Ascending	0.398	2.755	10.5	0.025	0.197	0.516
	Descending	0.531	2.264	8.573	-0.023	-0.169	-0.217
	ForwardStop	0.432	2.365	9.399	-0.024	-0.066	0.026
0.1	Ascending	0.392	2.566	9.515	0.005	0.101	0.302
	Descending	0.557	2.282	8.440	-0.028	-0.196	-0.258
50/50 mixture of Weibull(0.45, 1) and GPD(1.419845, 0.2) with $u = 0.4428726$							
	ForwardStop	0.292	1.328	4.082	-0.031	0.063	0.245
0.01	Ascending	0.302	1.768	6.352	-0.028	0.319	0.878
	Descending	0.374	1.283	3.729	-0.045	-0.050	0.037
	ForwardStop	0.306	1.186	3.370	-0.033	-0.070	-0.067
0.05	Ascending	0.295	1.417	4.509	-0.040	0.123	0.409
	Descending	0.428	1.236	3.128	-0.038	-0.203	-0.327
	ForwardStop	0.333	1.168	3.251	-0.037	-0.114	-0.150
0.1	Ascending	0.298	1.305	3.944	-0.044	0.044	0.219
	Descending	0.442	1.247	3.077	-0.029	-0.211	-0.367

The first task is to find a suitable threshold, such that the GPD fits the observed PM2.5 exceedances over this threshold. In the threshold selection process, we combine three different approaches to find the optimal threshold, including ForwardStop, ascending and descending goodness-of-fit tests. Second, derive the exceedances over the chosen threshold from the above three threshold selection methods. Then the PM2.5 upper record statistics can be obtained. Third,

Table 4. Threshold selection and upper records prediction

Threshold	Upper records	Method	$X_{U(s+1)}$	$X_{U(s+2)}$
10	11, 18, 80, 86, 101, 133, 162, 165, 183, 241, 242, 254, 365, 393, 430	LS	480	535
		WNLSM	479	532
188	241, 242, 254, 365, 393, 430	LS	467	509
		WNLSM	468	511

Table 3. Threshold selection and parameter estimation

	Threshold	Parameter	LS	WNLSM
ForwardStop	10	ξ	0.080	0.077
		σ	13.4	13.8
Ascending	10	ξ	0.080	0.077
		σ	13.4	13.8
Descending	188	ξ	0.114	0.136
		σ	26.9	25.1

estimate the parameters of (ξ, σ) with the upper record values using the LS and WNLSM methods. The results of threshold selection and parameter estimation are presented in Table 3.

After parameter estimation of the GPD, the main interest is the further upper record prediction. The most well-known predictor is the best linear unbiased predictor (BLUP) (see Ahsanullah (1995) subsection 4.4). Theorem 2.1. is applied to obtain the BLUP of the future upper record values $X_{U(s+1)}$ and $X_{U(s+2)}$ based on the first s observed records. Table 4 displays the results of threshold selection and upper records prediction. Notably, the future upper records are similar based on the LS and WNLSM techniques. These values can serve as references for environmental agencies and Beijing residents.

5. CONCLUSION

In practice, the shape parameter of the GPD is typically unknown; however, the existing estimation methods all assume that the shape parameter is known. We proposed a new procedure to estimate all the parameters of the GPD for upper record values. The new estimator appropriately addresses a caveat of the GPD shape parameter estimation. Our method is adapted from the moment method and nonlinear weighted least squares theory in the optimization procedure. It provides analytical solutions and is computationally efficient and stable. Using extensive numerical studies, we found that the performance of the proposed WNLSM estimator is highly competitive in estimating the parameters of the GPD. In addition, to investigate the performance of the ForwardStop rule, ascending and descending goodness-of-fit tests, various simulation studies were conducted. The results showed that no testing procedure outperformed in all conditions. However, ForwardStop is more stable and less sensitive to other two competing threshold selection procedures.

From a practical perspective, the proposed method was applied to the Beijing daily PM2.5 data. In this application, we used the ForwardStop rule and ascending and descending goodness-of-fit tests to select thresholds and upper records. Based on the records, future upper record statistics were predicted by a BLUP. These results serve as references for environmental agencies and Beijing residents. In summary, the proposed method was successfully applied to environmental data, and important knowledge with respect to the degree of air pollution was attained.

ACKNOWLEDGEMENTS

The authors are very grateful to the editor and anonymous reviewers for their insightful and constructive comments and suggestions.

FUNDING

The authors gratefully acknowledge the support of National Natural Science Foundation of China through Grant No. 11801019, Science and Technology Program of Beijing Education Commission through Grant No. KM201610005020.

Received 17 September 2018

REFERENCES

Ahsanullah M., Record statistics. *Nova Science Publishers, Inc*, 1995. MR1443904

Ahsanullah M., Record values–theory and applications. *University Press of America*, 2004. MR1370417

Arnold B.C., Balakrishnan N., Nagaraja H.N., Records. *Wiley*, New York, 1998. MR1628157

Bader B., Yan J., eva: Extreme value analysis with goodness-of-fit testing. R package version 0.1.2, 2015.

Bader B., Yan J., Zhang X., Automated threshold selection for extreme value analysis via ordered goodness-of-fit tests with adjustment for false discovery rate. *Ann. Appl. Stat.*, 12(1): 310–329, 2018.

Balakrishnan N., Chan P.S., On the normal record values and associated inference. *Statist. Probab. Lett.*, 39(1): 73–80, 1998.

Balakrishnan N., Ahsanullah M., Recurrence relations for single and product moments of record values from generalized Pareto distribution. *Commun. Statist. Theor-Meth.*, 23(10): 2841–2852, 1994.

Balkema A.A., De Haan L., Residual life time at great age. *Ann. Probab.*, 2(5): 792–804, 1974.

Chandler K.N., The distribution and frequency of record values. *J. Roy. Statist. Soc. Ser. B*, 14(2): 220–228, 1952.

Choulakian V., Stephens M.A., Goodness-of-fit tests for the generalized Pareto distribution. *Technometrics*, 43(4): 478–484, 2001.

Coles S., An introduction to statistical modeling of extreme values, 1st ed. *Springer*, Berlin, 2001.

Coles S.G., Tawn J.A., A Bayesian analysis of extreme rainfall data. *J. Roy. Statist. Soc. Ser. C*, 45(4): 463–478, 1996.
Cramer E., Naehrig G., Laplace record data. *J. Statist. Plann. Inference*, 142(7): 2179–2189, 2012.
Davison A.C., Smith R.L., Models for exceedances over high thresholds. *J. Roy. Statist. Soc. Ser. B*, 52(3): 393–442, 1990.
Dress H., De Haan L., Resnick S., How to make a hill plot. *Ann. Statist.*, 28(1): 254–274, 2000.
Embrechts P., Klüppelberg C., Mikosch T., Modelling extremal events for insurance and finance. *Springer-Verlag, Berlin Heidelberg*, 1997.
G'Sell M.G., Wager S., Chouldechova A., Tibshirani R., Sequential selection procedures and false discovery rate control. *J. Roy. Statist. Soc. Ser. B*, 78(2): 423–444, 2016.
Langousis A., Mamalakis A., Puliga M, Deidda R., Threshold detection for the generalized Pareto distribution: Review of representative methods and application to the NOAA NCDC daily rainfall database. *Water Resour. Res.*, 52(4): 2659–2681, 2016.
Nelder J.A., Mead R., A simplex method for function minimization. *Comput. J.*, 7(4): 308–313, 1965.
Park M.H., Kim J.H.T., Estimating extreme tail risk measures with generalized Pareto distribution. *Comput. Statist. Data Anal.*, 98: 91–104, 2016.
Pickands J., Statistical inference using extreme order statistics. *Ann. Statist.*, 3(1): 119–131, 1975.
Raqab M.Z., Inferences for generalized exponential distribution based on record statistics. *J. Statist. Plann. Inference*, 104(2): 339–350, 2002.
Scarrott C., MacDonald A., A review of extreme value threshold estimation and uncertainty quantification. *Revstat-Stat. J.*, 10(1): 33–60, 2012.
Shahab M.A., Al-Dayian G.R., El-Beltagi S.H., Analysis of the time intervals between the academic degrees of university faculty members and graduate assistants by using the record values. *J. Fac. Commer. Rev.*, 18: 451–480, 2001.
Soliman A., Amin E. A., Abd-El Aziz A. A., Estimation and prediction from inverse Rayleigh distribution based on lower record values. *Appl. Math. Sci.*, 4(62): 3057–3066, 2010.
Sultan K.S., Al-Dayian G.R., Mohammad H.H., Estimation and prediction from gamma distribution based on record values. *Comput. Statist. Data Anal.*, 52(3): 1430–1440, 2008.
Sultan K.S., Moshref M.E., Record values from generalized Pareto distribution and associated inference. *Metrika*, 51(2): 105–116, 2000.
Wang B.X., Ye Z.S., Inference on the Weibull distribution based on record values. *Comput. Statist. Data Anal.*, 83: 26–36, 2015.

Xu Zhao
College of Applied Science
Beijing University of Technology
Beijing, 100124
China
E-mail address: zhaox@bjut.edu.cn

Xueyan Geng
College of Applied Science
Beijing University of Technology
Beijing, 100124
China
E-mail address: 971748452@qq.com

Weihu Cheng
College of Applied Science
Beijing University of Technology
Beijing, 100124
China
E-mail address: chengweihu@bjut.edu.cn

Pengyue Zhang
Department of Biomedical Informatics
College of Medicine
The Ohio State University
Columbus, Ohio
USA
E-mail address: zhangpe@imail.iu.edu